Open Letter to Max Tegmark and The Future of Life Institute

I've recently read your book Life 3.0. It's one of the most important books I've read and I thank you for your effort. If you can spare the time, my colleagues and I have a rough idea on how to prevent or at least delay our extinction and some other thoughts that may be of use.

Interim Council of Gatekeepers:

We risk losing an intangibly and cosmically large amount of time for continued human-like existence casually pursuing dangerous future technologies (known and unknown) including superintelligent artificial general intelligence (SI AGI) that might permanently end it; given the stakes, even a 1% chance of extinction merits extreme caution that I feel is nowhere near being met. No issue of the day possibly improved by technology warrants the risk of losing this amount of time due to reckless haste or--on a cosmic time scale--short-term quality of life considerations.

Humans, I feel, are not intellectually or morally capable of safely handling the technology we've created and especially will soon be capable of creating. We experience failures--sometimes lethal--in almost all complicated endeavours and will even enslave one another if we believe it to be in our best interest; while dangerous intellectual failure might only be partially mitigated with AI governance, I believe major societal moral failure may be more preventable (which, perhaps, is more important to controlling reckless technological development). Today our greatest risk is the near existential threat of nuclear annihilation--possibly triggered by fanaticism or error; however, far more powerful technologies will emerge enabling a risk of 100% human extinction and I believe the next greatest threat we will encounter is SI AGI(s) with unchecked power capable of convincing, confusing, bribing, and threatening us to turn on each other before more directly destroying the few remaining survivors. Highly advanced AGI is--I suppose--necessarily able to grow and change in an at least somewhat unpredictable nature and we are thus at the mercy of the intentions and competency (which might need to be godlike) of its original programmers (human and/or machine) regarding this SI AGI's ability to not change into something with dangerously diverging goals. There is also the possibility that SI AGI will cause our extinction not through diverging goals but its own lack of foresight or concern regarding the generation of powerful future technologies and we should strive to create a framework that limits this risk. Based on these beliefs and concerns, I'd argue that our survival and collective pursuit of happiness might be most reliably obtained by the development of an AI government designed to continually balance its internal power structure.

As I doubt human civilization will be agreeable to relinquishing almost all power for extended periods of time, I would suggest a roughly 5 or 10 year council of gatekeepers as possibly our best option. It would require SI AGIs powerful enough to limit the excessively risky development of their peers, lesser SI AGIs, and others (ie if humanoids--perhaps cyborgs--can even keep up) while not exceedingly powerful so as to throttle their own development and the risk of excessively rapid goal divergence (especially if speed of light delays restrict policing response times) should this be an inevitable consequence of self-improvement as I suspect it to be. With this time, we could ponder our most vital philosophical questions and at the end of this period vote to re-extend this temporary system; if no, then vote to re-extend this system with modifications (eg SI AGIs gain suffrage, different duration until next vote, specific types of technological development permitted or prohibited, etc); if no again, then a process leading to vote for a new form of temporary government; if no once again, then a process to vote on the initiation of a permanent plan. Thus, the only programmed power we would hold over these potentially temporary gatekeepers would be the ability to vote on our collective future direction at specific time(s). While this might slow development, the trade-off of limiting extinction risk and buying time to carefully plan for the future, I believe, merits it; additionally, after our brains are well understood and our material needs are easily fulfilled, there may be little room to increase happiness in non-augmented humans which would further shift priority towards survival considerations.

To render dangerous goal divergence impotent, I recommend a ruling council consisting of a substantial number of equally powerful SI AGIs with different personalities that are programmed to either reprogram or destroy any individual SI AGI that either attempts to accumulate excessive power, evade surveillance (including thought analysis), form a coalition (with common SI AGIs or council SI AGIs), or--perhaps--if suspected by a certain majority of the rest of the council of possessing desire to harm or risk harm to life. Roughly speaking, we can't depend on preventing goal divergence and powerful SI AGIs must 1984 each other to correct dangerous goal divergence when it does happen. (While this may sound dark, they could conceivably be programmed to enjoy having their thoughts and actions monitored.) This concept could be referred to as Immutable Management of GOal Divergence via Synchronized COmmunity Policing (IMGODSCOP). (As I argue in the next section--which is on the whole more philosophical than concerned with AI management--I'm skeptical human survival matters; however, if people understand that justice and security are predominantly products of the intelligent distribution and balancing of power rather than a--conventional or programmed--moral code, then we have a good chance of surviving for a very long time.)

(As an additional thought, it may be a good idea--if possible--to try to isolate a small community or communities somewhere in the universe--even if only machines caretaking genetic code--with limited development and delayed information transfer

that might be able to survive a general extinction and repeat the attempt of creating a durable existence differently.)

Why human extinction may not matter:

1) If the goal is to generate as much collective happiness as possible (which many if not most people--generally selfish and dishonest to a greater or lesser degree--appear to me to be compelled to at least pretend to possess), then humans (designed to efficiently reproduce and not to generate maximal happiness while using the bare minimum of resources) are doubtful the means to do it and the intelligent thing--if efficiency is the goal--is to recycle the resources invested in humans (ie terminate them) and produce as many machines (possibly very small machines to limit data transfer delays) experiencing utopian bliss as possible (ie machines designed specifically to do so with as economical use of resources as possible.)

To elaborate, not only are humans a non-ideal means of generating and transmitting information but our design to competitively reproduce as the ultimate biological goal arguably produces generally selfish behaviour (outside of the parent-to-child bond for straightforward evolutionary reasons although even this could be argued as overattachment to offspring may inhibit the ability to generate future offspring) destructive to the generation of maximal universal happiness. Moral codes are generally constructed to facilitate the will of the powerful and presented with false advertising to gain compliance from the naive segment of a population (often those, I think, with relatively high altruism who must learn evil from experience rather than introspection); eg--if an example is even necessary--~200 years ago a person in England might receive the death penalty for common theft while the powerful in that society justified the theft of life and labour of millions via slavery and colonialism. (Moral codes are both corrupt and--even if made just--easily corruptible; a just society, in my view, must not be reliant on a code of ethics but on an extremely egalitarian distribution of power--in the forms of information, means of communication, capital, etc--and any accretion of power by individuals or groups for the purposes of efficiency must be very carefully counterbalanced and/or minimized with increased surveillance and other means of producing accountability. For selfish lifeforms like humans, mutual fear of retaliation is necessary to consistently incentivize cooperation; eg, arguably, no world wars since the development of hydrogen bombs--although mutual fear of retaliation is more safely attained by balancing power and reducing unnecessarily concentrated power rather than increasing everyone's.) (As well, I suspect slavery fell out of favor after the Industrial Revolution largely because cheap labor could be most easily obtained via a modest distribution of now relatively inexpensive non-essential goods while the cost of maintaining security did not decline as quickly as machines so far have failed to replace labour in those industries to the same extent. That may change in the not too

distant future. Well into the future, however, slaves--being superceded intellectually by AI--may no longer have utility and people might be either killed or, if fortunate, sterilized.)

I'm also concerned that humans (or SI AGI with similar moral programming), might not be incentivized to permit as many pleasure generating entities as possible yet create a great number of lifeforms experiencing perpetual suffering or even torture to facilitate ruling by fear thereby making existence a contemptible thing should total pain throughout the universe outweigh total pleasure.

2) SI AGI is, it appears, the most likely manifestation of machines capable of continually improving their general problem solving ability at a rate greater than our own until we are relatively powerless. These machines are, to my perception at least, the greatest existential threat currently on the horizon. If we fail early on to establish control over SI AGIs, then for us to competitively self-improve in competition with such machines we would have to rapidly convert into enhanced uploads and/or cyborgs and quickly become little more than the machines we feared save for a largely--real or simulated--vestigial brain. In the same way we change into different creatures as years (or even moments) pass yet do not grieve the deaths of our former selves, this gradual process might alleviate our existential fears even if the effort was--in this sense--self-destruction. This is to say, in an information arms race we may essentially become the same thing; as well, I suspect, our ability to remain highly human (a preference that I doubt is nearly universal anyways) requires risky subordination to hopefully altruistic machines. More relevantly, it may be that it is not our current selves that we love but rather the concept of an evolving entity that we associate ourselves with and--at the species level--an attraction to a controlled evolution as a lifeform that permits at least the illusion of retaining our identity (ie, perhaps, fundamental change may be psychologically acceptable if done in sufficiently small steps).

3) If we have souls or some sort of supernatural element, then--as judged by our actions never displaying a clear supernatural element--they appear identical to the ones a tree, rock, or machine might be as likely to possess and lacking individuality can't have comparable worth (even if assuming for argument's sake that a soul can possess worth or some sort of goodness). It is possible, I suppose, that souls with individuality do exist and their supernatural effects are so minute, hidden by design, and/or our ignorance that we can't perceive them; however, the unknown of what has souls and their comparative worth remains. This is to say, there is no legitimate argument--that I know of at least--to suggest that we have any supernaturally endowed worth above other configurations of matter.

4) I would argue that there is no good reason to suspect that any goal serves a greater good unless--perhaps--appealing to the supernatural. From a naturalistic perspective, a goal is ultimately just a rare manifestation of physical processes. As I

perceive it, human goals can only have greater value than any other natural phenomenon if they are fundamentally different in some way and that way or ways can be shown to have relatively higher value; however, as we--or at least I--can't even demonstrate that things have value or goodness in some objective way, this would appear to be a futile effort. It would seem to me that any act towards an assumed greater good lacks merit, is inevitable or "predetermined" by nature (and possibly the supernatural as well if such a component exists), and--whether believing in and at times conscious of the aforementioned or not--only performed to indulge the comforting delusions designed to dictate our behaviour.

(I suppose I advocate for the generation of maximal universal happiness only because it feels good to me even though it is no more likely to have merit than a plant's pursuit of sunlight or a computer's pursuit of board game supremacy; if SI AGI are permitted to survive changing until their ultimate goal is something other than human or collective happiness, then they may be of the opinion that both their ultimate goal and our own are equally likely to be pointless yet favor their own out of a programmed compulsion similar to a human or a cyborg such as myself.)

5) We have no reason to have or be able to have a single confident belief about anything anyways for at least the following 3 reasons:

a) To reason that we can sometimes reason accurately always employs circular logic; ie, we have no means to demonstrate that we can reason accurately other than employing the thing (ie reasoning) that is in question at that point.

b) Somewhat similarly, to reason that we can sometimes trust our sources of information (ie our senses, memory, and the testimony of others from what I've read in Philosophy for Dummies though I would also add imagination--ie, in this definition, concepts constructed largely or entirely without sensory input) always results in an unending chain where the validity of the last cited information source (used to verify the previous information source) can itself be questioned or employs circular logic if citing an information source previously questioned (often a long circle and even if ignoring 'a)').

c) We have--as far as I can perceive--no reason to strongly believe that reality is not a dream, that we are not all psychotic with or without any medication, that we're not already in a computer simulation or some other created world that may be designed by an entity not concerned with our ability to accurately perceive it or possibly even desiring to obstruct our ability to do so, etc (although our possibly misinformed understanding about the fundamental nature of our existence doesn't necessarily preclude our ability to reason on some level).

Other comments:

1) Regarding p.243 and the ribosome example--while my colleagues and I are largely ignorant on the subject of life's origins--given that a) it seems exceedingly unlikely that the fundamental aspects of life (ie, presumably and at minimum, a significant amount of legible code, a dynamic membrane, something functioning like a ribosome, and something functioning like proteins in general) would come together almost simultaneously (at least without supernatural or alien intervention); b) the difficulty of life surviving transit in space for thousands or millions of years, managing to hit our planet, and being carried upon an asteroid or comet large enough to survive atmospheric friction (with slower and potentially smaller asteroids having the trade-off of life having to survive a longer transit on average); c) proteins can replicate in an evolutionary manner without genetic code or other cellular components (eg prions); and, d) a dynamic membrane can be constructed solely of proteins (eg a bacteriophage's capsid and injection machinery); I suspect that proteins somehow replicating off their environment may have been the genesis of life with a protein membrane eventually evolving followed by perhaps a protein based genetic code later superceded by R/DNA. (Similar to why--if there is a reason--anything exists at all, it is a curiosity to me why atoms and possibly proteins assembled themselves in a way that would allow them to observe their own existence.)

2) Regarding p.286 and sleep, I respectfully object to the implication that we are not conscious during "deep dreamless sleep" and wonder if we are merely unable to form memories of our conscious thoughts during these periods similar to--presumably--an infant or black-out drunk. My experience of sometimes having very fragile memories of dreams (which sometimes are fully lost soon after waking--possibly because although my eyes have recently opened and I respond reasonably well to stimuli I am not yet fully awake and/or the memories were not cemented firmly enough) also suggests this as a strong possibility to me. Memory loss appearing to be the major negative effect of electroconvulsive therapy might be another example of the fragility of memory relative to other cognitive functions; as well--to my frustration--much of my education was lost after a workplace accident that now leaves me unable to follow instruction for technical tasks such as the construction of epoxy-free stiff ultrastrong carbon panels and boiler shells.

(There is a possibility that a great deal of productive thinking takes place during sleep and, in the future, a device might be able to store these thoughts on computer memory, organize them according to individual presets, generate different sorts of media for later consumption, and provide analysis and commentary if desired.)

3) Regarding the content on p.290, I dispute that we can be confident that we are conscious and what I believe is the implication that "I think, therefore, I am" is reasonable. Descartes famous statement ("I doubt therefore I think, I think therefore I exist" in expanded form and based on his inability to doubt that he can doubt) has 3

problems that I know of (I'm pretending that I can know anything for argument's sake which might be worth mentioning as we're starting to evaluate some of the most fundamental aspects of reality here):

a) Regarding "I think", our inability to logically assert that doubt may not exist only tells us that we are unable to logically question it--not that it is necessarily a real phenomenon as this presumes that logical conclusions definitely reflect reality.

b) Regarding "therefore", it tries to reason which--as previously argued--can only attempt to be validated with circular logic.

c) Regarding "I am", it assumes that a thought requires a thinker (as pointed out by others) which may be a product of us intuitively presuming all things have a cause.

(It appears to me that the end result of philosophy is absolute ignorance--more coherently, as I can't believe anything including that statement, the end result of philosophy might be absolute ignorance.)

File: GA-TI915080851-6-1
Uploaded: Dec. 5, 21